DBN Based Joint Dialogue Act Recognition of Multiparty Meetings

DBN BASED JOINT DIALOGUE ACT RECOGNITION OF MULTIPARTY MEETINGS

Alfred Dielmann and Steve Renals

Centre for Speech Technology Research
University of Edinburgh, Edinburgh EH8 9LW, UK
Email:{a.dielmann,s.renals}@ed.ac.uk

ABSTRACT

Joint Dialogue Act segmentation and classification of the new AMI meeting corpus has been performed through an integrated framework based on a switching dynamic Bayesian network and a set of continuous features and language models. The recognition process is based on a dictionary of 15 DA classes tailored for group decision-making. Experimental results show that a novel interpolated Factored Language Model results in a low error rate on the automatic segmentation task, and thus good recognition results can be achieved on AMI multiparty conversational speech.

Index Terms— DA, DBN, Interpolated FLM, AMI

1. INTRODUCTION

Dialogue acts (DAs) represent the function that utterances serve in a conversation and aim to capture the intentions of a single speaker. A DA annotation scheme provides a set of disjoint classes that may be used to label every possible conversational act. We are interested in the automatic recognition of DAs in the context of multiparty conversational speech, since they can highlight precious facets of the discourse structure and provide a valuable input for other research areas, like: topic detection, automatic summarisation, hotspot detection, etc . DA recognition consists of two steps, which may be performed sequentially or jointly: (1) subdividing the transcribed text in DA segments (DA segmentation) [1]; (2) classifying each segment as one of the DA classes defined by the annotation scheme (DA tagging or DA classification). The aim is to extract automatically a sequence of DA units which is similar to the reference DA sequence provided by human annotators. In this work we focus on joint DA recognition using trainable statistical models based on dynamic Bayesian networks (DBNs) and factored language models (FLMs).

Previously we focused on DA recognition using the ICSI meeting corpus [2], using 5 broad DA categories (statements, questions, disruptions, fillers and backchannel). In this paper we focus on the AMI meeting corpus (section 2) and its annotation scheme oriented towards decision-making, with 15 DA classes. Since this corpus has been collected and released recently this is the first work reporting on automatic DA recognition applied to the AMI meetings. The AMI DA annotation scheme not only has three times more DA classes than ICSI (exacerbating data sparsity) but also includes more abstract speaker intentions such as reassuring the group and commenting on previous discussions. This results in a more challenging task.

The main contributions of this paper are some new approaches to automatic DA segmentation and classification based on interpolated FLMs and a hybrid DBN infrastructure. The hybrid DBN (section 3.3) helps to combine the accurate DA segmentation that is possible with the interpolated FLM together with the discriminative properties of conventional FLMs. We present experimental results for DA tagging, segmentation and recognition using these techniques, applied to the AMI meetings corpus.

This work was partly supported by the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6–506811, publication AMI–198).

2. THE AMI MEETINGS CORPUS

The AMI Meeting Corpus [3] is a multimodal data set comprising 100 hours of multiparty meeting recordings. This corpus consists of two homogeneous sub-sets of data: 138 "scenario meetings" in which the participants play different roles in a design team, taking a design project from kick-off to completion, and 34 (about one-third of the entire data corpus) naturally occurring "non-scenario meetings". This corpus was collected at three meeting rooms instrumented with a comprehensive set of synchronised recording devices including close-talking and distant microphones, individual and room-view video cameras, whiteboard capture, and digital pens. The entire corpus [1] has been manually annotated with orthographic transcriptions, and several different phenomena (such as dialogue acts, hand movements, head movements, named entities, and topics) have been annotated for most of the corpus.

Since the annotation of DAs has been completed only for the "scenario data-set", all the experiments reported in this paper have been performed on this subset only. The scenario subset has been subdivided into training (71% of the available data, around 0.4 millions of words), development (14.5%) and test sets (14.5%). The DA annotation scheme for the AMI corpus is based on 15 DA classes tailored for group decision-making. Each DA unit highlights a single speaker intention and classifies it into one of the following 6 broad DA categories and 15 specialised DA classes: information exchange and questions (*inform* and *elicit inform*), action that an individual or group might take (*suggest*, *offer* and *elicit offer or suggestion*), comments on previous discussion (*assess*, *elicit assessment*, *comment about understanding* and *elicit comment understanding*) , actions targeted on group's social functioning (*be positive* and *be negative*), part of the transcriptions that are not real DAs (*backchannel*, *stall* and *fragment*), other speaker intentions not covered by the previous classes (*other*). We are concerned with the automatic recognition of the 15 more specific classes, since they provide deeper insights on speaker intentions, compared with the 6 broad categories. The distribution of the 15 DA classes across the corpus is shown in table 1. Not surprisingly information exchange (*inform* is the most frequent class) is the prevailing task of these product-design oriented meetings. Backchannel and fragment units (second and fourth rows

[1] Both raw recordings and manual annotations are freely available and can be downloaded from http://corpus.amiproject.org/.

Dialogue Act class	% DAs	Dialogue Act class	% DAs
Inform	26.57 %	Other	1.78 %
Backchannel	17.60 %	Be positive	1.75 %
Assess	16.68 %	Elicit assessment	1.71 %
Fragment	12.98 %	Offer	1.18 %
Suggest	7.47 %	Elicit offer or sugg.	0.53 %
Stall	6.33 %	Elicit comment und.	0.16 %
Elicit inform	3.37 %	Be negative	0.07 %
Comment underst.	1.81 %		

Table 1. DA classes distribution (% of the total number of DA units).

of table 1) are highly characteristic of natural conversations. Note that the top five classes of table 1 account for more than 80% of the total DA units, and intentions like *be negative* and *elicit comment understanding* are extremely rare.

3. AUTOMATIC JOINT DIALOGUE ACT RECOGNITION

Our system for the automatic joint DA segmentation and classification relies on a supervised statistical approach, employing two language models, a Gaussian mixture model (GMM), and several conditional probability tables (CPTs), with all parameters estimated from training data.

The system core (section 3.3) consists of a switching DBN, which combines several subsystems and coordinates the entire recognition process. An interpolated factored language model (FLM) is used to relate sequences of transcribed words to their corresponding DA labels (section 3.2). Trigram statistics are used to model the sequence of DA units: the "discourse model". Prosodic information, such as pitch and energy, is extracted from the audio signals (section 3.1) and used for the segmentation sub task. The DBN framework that we use operates only on discrete and quantised states, therefore all these continuous word related features are mapped into discrete states through a GMM. The number of Gaussian mixture components is automatically learned during training. For practical reasons the overall learning process is subdivided into two steps: both FLM and discourse model are trained independently before being embedded into the DBN infrastructure, while the GMM and all the CPTs associated to the DBN topology are trained later in a second phase. During testing the whole system processes unseen meeting data providing labeled DA segments as output. If segmentation alone is required, it is possible to discard the DA labeling information. Conversely a known DA segmentation can be used to override the automatic segmentation process, thus forcing the system to operate as a standard DA tagger.

3.1. Continuous word related features

Six normalised continuous features, extracted from the audio signals and from orthographic transcriptions, were associated with each processed word: F_0 *mean*, F_0 *variance*, *RMS energy*, *word length*, *pause duration* and *word relevance*. Figure 1 shows the feature extraction process. Orthographic transcriptions are automatically aligned against the raw audio recordings in order to estimate word temporal boundaries [3]. A precise temporal location is needed to improve feature quality: accounting for the exact signal segment referring to a given word w_i helps to reduce estimation errors and provide more discriminative features. Intra-word mean and variance of the fundamental frequency F_0 is estimated for each word using the ESPS pitch tracking algorithm and are made speaker independent by normalising them against the average baseline pitch for that speaker. The root mean square energy is computed for each word w_i, and then normalised both against the average channel energy and against the mean energy of all the other occurrences of that word w_i. Word length is simply obtained as the ratio between the uttered word length, estimated by forced alignment, and the average typical duration for that word w_i. Pause duration refers to the length of the silence segment existing between each word w_i and the following one w_{i+1}. This feature despite its simplicity is usually well correlated with sentence breaks and often offers a good prior to detect DA boundaries [4]. Pause related features are even more effective if taken in conjunction with a richer set of prosodic related features.

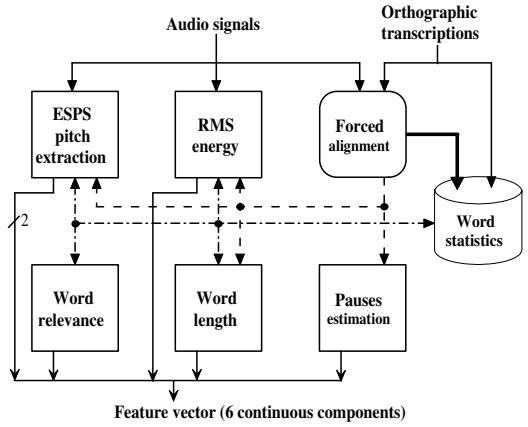

Fig. 1. Data flowchart of the feature extraction process.

3.2. Interpolated Factored Language Models

DA taggers are often based on statistical language models (LMs) which are used to discriminate between multiple labeling hypothesis. DA segmentation could also take advantage from LMs since some words or constructs are good cues to highlight DA boundaries (e.g., the wh-words: when, what, who, ... which could often predict the beginning of direct questions). Moreover in a joint DA recognition task both DA boundaries and labels are unknown and a LM can be used to select the optimal labeled segmentation between all the possible hypotheses. In practice, pruning is essential to reduce the computational effort.

Several language models have been applied to DA related tasks, including hidden event language models [5], class based language models [6] and factored language models (FLMs) [7]. FLMs are the natural extension of conventional LMs where each word w_t is replaced by a bundle of factors: $\mathbf{v}_t \equiv \{v_t^0, v_t^1, \ldots, v_t^k\}$. Part of speech tags, morphological classes, word position, and token words themselves are all typical examples of factors. The goal of conventional LMs is to factorise the joint distribution $p(w_1, w_2, \ldots, w_n)$ as a chain product of conditional probabilities in the form $p(w_t \mid w_{t-1}, ..., w_{t-n})$; similarly with FLMs the joint distribution $p(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ will be factorised as the product of terms like $p(\mathbf{v}_t \mid \mathbf{v}_{t-1}, ..., \mathbf{v}_{t-n})$. Often $v_t^0 = w_t$ since we are usually interested in predicting the current word w_t given the previous factor history. As usual smoothing techniques are necessary during parameter estimation. Therefore in FLMs there are two degrees of freedom: choice of the optimal (in the context of a well defined task) factor set, and selection of which factors need to be discarded during every backoff step (FLM topology). Note that, even if not used in this work, FLMs are often associated with

the concept of generalised parallel backoff [7]: capability to follow multiple concurrent backoff strategies.

Our FLM has been chosen to maximise DA tagging accuracy and is based on three factors: words w_t, DAs d_t, and the relative position of each word in the DA unit n_t. Each DA unit is composed by a sequence of words w_t, w_{t-1}, ... w_{t-k} with the same DA label $d_t = d_{t-1} = \ldots = d_{t-k}$. n_t takes in account the relative position of each word w_t in the DA unit, thus each DA segment has been subdivided in blocks of five words, with n been incremented after every block of 5 words. Since the vast majority of the DA units contains fewer than 75 words, the position factor is constrained to have a maximum $n = 15$. The proposed model is defined by a product of conditional probabilities in the form $p(w_t \mid w_{t-1}, n_t, d_t)$. w_{t-1} is dropped during the first backoff step leading to $p(w_t \mid n_t, d_t)$. When a further backoff is required, DA labels are the new dropped factor, leading to $p(w_t \mid n_t)$. Both backoff steps are smoothed using Kneser-Ney discounting.

Multiple FLMs sharing the same topology can be interpolated into a single FLM. This principle here is applied in order to train a FLM from multiple weighted data sources, in this case the ICSI meeting corpus and the FISHER corpus of conversational telephone speech [2]. Although FISHER lacks DA annotation and the ICSI DA categories are incompatible with the AMI DA annotation scheme, it is possible to exploit this data by enriching our FLM and improving DA segmentation. This is achieved by duplicating the content of both FISHER (10.62M words) and ICSI (0.74M words) corpora 15 times, labeling each sentence with all the 15 available DA classes, training the FLMs for the three corpora and finally interpolating them. The resulting model has a richer vocabulary, and a richer set of n-grams, since word sequences absent from the AMI training data set are now directly represented by the LM.

3.3. Dynamic Bayesian Network based model

The core of our DA recognition framework consists in a switching DBN [8] (figure 2) which integrates two FLMs, a discourse language model, and a GMM (implemented using GMTK [9]). The model alternates between two operating conditions implemented as two model topologies. The most frequent one (*intra-DA* topology depicted in figure 2-A) refers to the general case where a sequence of words is being processed as part of a single DA unit. When the end of a DA unit is detected (through the node E) this graphical model will switch to the second topology (*inter-DA* topology in figure 2-B). Both of the topologies are likely to be evaluated at every frame enabling DA boundaries to be placed where they are most likely.

The accumulated probability of the transcribed word sequence W_{t-k}, ..., W_{t-2}, W_{t-1}, W_t will be estimated by the FLM (dotted arc between W_{t-1} and W_t of figure 2-A). Note that a second FLM (gray dotted arcs) will be introduced when two FLM with complementary qualities are available, leading to a hybrid approach. Both FLMs $p(W_t \mid W_{t-1}, N_t, DA_t^0)$ also depend on DA labels (arc between DA_t^0 and W_t) and on the relative word position (node N_t). C_t is a bounded counter: from 0 to 4 and back to 0, used by N_t to split the transcribed text into blocks of five words. Since the entire model is based on the assumption that DA units generate sequences of words, this could be referred to as a generative approach. If DA_t^0 represents the currently hypothesized DA label, DA_t^1 and DA_t^2 contain the DA recognition history and need to be update only across DA boundaries (*intra-DA* topology). The DA boundary detector node E_t (shared by both the topologies) depends on the recognition history, the current DA label,

[2]Linguistic Data Consortium Catalog: http://www.ldc.upenn.edu/Catalog

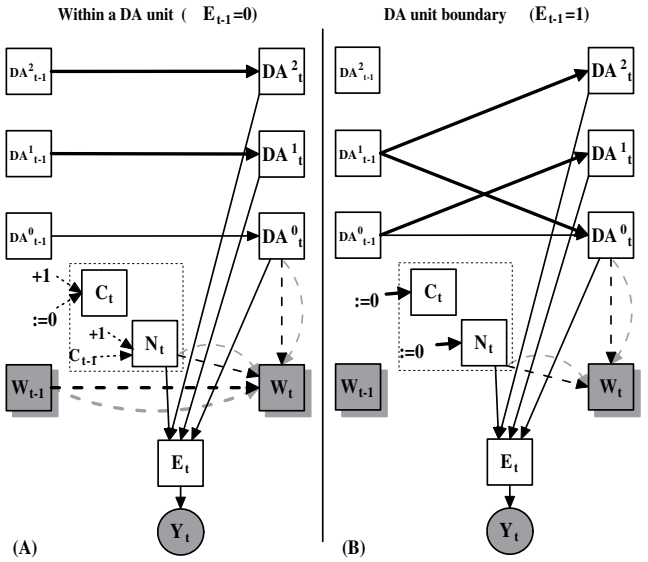

Fig. 2. Generative switching DBN for the automatic DA recognition tasks. The model alternates between two topologies: (A) inside a single DA segment (B) across two adjacent DA units, according to the boundary detector node E_{t-1}. Note that: unshaded square nodes represent hidden discrete states, shaded circles are continuous observable features, and dotted arcs show probabilistic dependencies implied by the FLM. Gray round dotted arcs highlight an optional additional FLM.

the relative word position N_t and is also used to generate the continuous word related features through a GMM. The aim of this variable is to forecast DA breaks and to switch between the two operating conditions. Note that we did not make use of any artificial parameter to influence the resulting segmentation insertion/deletion rate. When a DA boundary has been foreseen ($E_{t-1} = 1$) the *intra-DA* topology depicted in figure 2-B takes care of the model reinitialisation. In order to start the evaluation of a new DA unit: both C_t and N_t are set to zero, the trigram discourse model $p(DA_t^0 \mid DA_{t-1}^0, DA_{t-1}^1)$ generates a new set of DA labeling hypotheses, and the FLMs are forced to backoff to $p(W_t \mid N_t, D_t)$.

Node cardinalities are imposed by the variables they represent: $\mid DA_t^x \mid = 15$, $\mid C_t \mid = 5$, $\mid N_t \mid = 15$, $\mid E_t \mid = 2$ and W_t has as many states as the words contained in the FLMs. Note that the model depicted in figure 2 represents two generic BNs which are duplicated for $t > 0$. During $t = 0$ all the hidden variables need to be initialised properly: $DA_0^1 = DA_0^2 = 0$, $C_0 = 0$, and $N_0 = 0$.

4. EXPERIMENTS AND DISCUSSIONS

In this section we compare four different system configurations (using different FLMs) derived from the DA recognition framework outlined in the previous sections. The first one is based on an FLM trained only using the AMI training dataset (column *AMI* of table 2). The second setup (*iFLM1*) uses an FLM interpolated on AMI, ICSI and FISHER data, assigning to each contribution the same weight. The third configuration (*iFLM2*) makes use of a weighted interpolated FLM: data from the AMI training subset will account for the 58.5% of the FLM, 2.7% of the language model derives from ICSI meetings and 38.8% from the FISHER corpus. The last setup is a

Task	Metric	AMI	iFLM1	iFLM2	Hybrid
TAG.	100 - (% Corr.)	**40.85**	54.92	51.39	42.79
S	NIST-SU	70.68	45.97	**20.39**	25.63
E	DSER	78.02	65.79	**12.78**	17.05
G	STRICT	74.38	57.18	**28.52**	36.86
M.	BOUNDARY	10.76	7.00	**3.10**	3.90
R	NIST-SU	93.15	81.31	73.64	**71.32**
E	DER	85.48	82.47	57.02	**51.86**
C.	STRICT	83.16	75.40	64.42	**62.10**
	LENIENT	40.94	46.84	51.85	**42.21**

Table 2. DA Tagging, Segmentation and Recognition error rates (%) on the AMI meeting corpus using four different FLM setups.

Hybrid between the first and the third configuration, concurrently making use of two language models: the plain *AMI* FLM and the weighted interpolated FLM of *iFLM2*. Note that these experiments have been performed on reference orthographic transcriptions and not on speech recognition output.

If the reference DA segmentation is known (so nodes E_t instead of being hidden variables now contain discrete observable features), our DA recognition framework can be used as a simple DA tagger. Classification error rates for the three setups are reported in the top row of table 2. Both the interpolated FLMs have poorer classification performances if compared with a standard FLM. During interpolation many n-grams are discarded; those from the AMI data carry a true DA annotation, those from the ICSI and FISHER corpora have only fictitious non–discriminative DA labels. Therefore any loss of n-grams from the AMI subset will induce an inevitable degradation in the tagging accuracy. The test condition *iFLM2* represents a trade-off between the baseline tagging results achieved on a conventional FLM and the performance degradation induced by *iFLM1*. Note that all the DA tagging results are heavily influenced by the imbalanced distribution of the 15 DA classes (section 2). The percentage of wrongly classified units reaches 93.3% by drawing the DA classification by chance and 82.9% by taking in account the prior distribution of the DA classes. If every unit is classified as *Inform* (the most frequent class) the percentage of erroneously labeled units drops to 65.5%.

If performance evaluation of the DA tagging task is trivial, since we can easily measure the percentage of wrongly labeled DA units, both DA segmentation and recognition tasks leave space to many different evaluation metrics. In this work we have adopted the metrics used by Zimmermann et al. [10]: NIST "Sentence like Unit" derived metrics, strict, lenient, boundary based metrics, DA Error Rate (DER) and DA Segmentation Error Rate (DSER). Although the interpolated FLMs caused a degradation in the DA tagging accuracy, they led to a significant improvement in DA segmentation error rates. In particular *iFLM2* was able to demonstrate accurate segmentation on all the evaluation metrics. Recognition results also reflect the balance between segmentation and classification: interpolated FLMs perform better than the baseline *AMI* FLM thanks to the improved segmentation. Note that the LENIENT metric, which represents the percentage of erroneously classified words ignoring DA boundaries, shows that although *iFLM2* outperforms *iFLM1* in terms of DA tagging error rate, the DA units correctly classified by *iFLM2* usually contain fewer words than those from *iFLM1*. The baseline *AMI* FLM offers a good tagging error rate and the weighted interpolated FLM *iFLM2* provides an excellent segmentation. Combining both of them (*Hybrid* setup of table 2) in a DBN model with two concurrent language models, helps integrating these complementary strengths. The resulting framework has average tagging and segmentation performances, but also provides the best DA recognition output.

5. CONCLUSION

We have investigated the joint Dialogue Act recognition task on multiparty conversational speech, reporting the results on the AMI meeting corpus. Our approach used a modular framework integrating an interpolated Factored Language Model, a second FLM, a Gaussian mixture model, 6 continuous word related features, a 3-gram discourse model and a switching graphical model. In the resulting system every component is focused on both the DA segmentation and classification tasks: as the lexical content is responsible of the tagging but has also a severe impact on the segmentation, the 6 prosodic features are primarily responsible for the DA segmentation but the DA classification task can be facilitated by a good segmentation.

The obtained results not only offer a good baseline for the automatic recognition of DAs on AMI meetings, but also show the advantages of interpolating multiple factored language models trained on similar corpora. Such interpolation results in a reduced tagging accuracy (a loss of about 10.5%), but a considerable improvement in segmentation accuracy, with the number of DA segmentation errors being halved, leading to a NIST-SU segmentation error rate of 20.4%.

6. REFERENCES

[1] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, D. Jurafsky, P. Taylor, R. Martin, C. Van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, 2000.

[2] A. Dielmann and S. Renals, "Multistream recognition of dialogue acts in meetings," *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–06)*, 2007.

[3] J. Carletta et al., "The AMI meeting corpus: A preannouncement," *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2006.

[4] M. Zimmermann, A. Stolcke, and E. Shriberg, "Joint segmentation and classification of Dialog Acts in multiparty meetings," *Proc. IEEE ICASSP*, May 2006.

[5] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," *Proc. Interspeech - ICSLP, Philadelphia*, Oct. 1996.

[6] B. Suhm and A. Waibel, "Towards better language models for spontaneous speech," *Proc. Interspeech - ICSLP, Yokohama*, Sept. 1994.

[7] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," *Proceedings of HLT/NAACL 2003*, May 2003.

[8] J.A. Bilmes, "Dynamic Bayesian multinets," *Proc. Int. Conf. on Uncertainty in Artificial Intelligence*, 2000.

[9] J. Bilmes and G. Zweig, "The Graphical Model ToolKit: an open source software system for speech and time-series processing," *Proc. IEEE ICASSP*, Jun. 2002.

[10] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," *Proc. Multimodal Interaction and Related Machine Learning Algorithms Workshop (MLMI–05)*, 2006.